



Choice of controls in case-control studies

Pontus Naucner, MD PhD

Infectious Disease Unit, Dept. of Medicine, Solna, Karolinska
Institutet

“I would trust only six people in the world to do a proper case-control study.”

David L. Sackett

ESCMID Online Lecture Library © by author

”Problems of case-control studies are that they are too small, poorly designed, improperly analysed or overly interpreted”

”Medical science and public health would be well served by fewer, larger case-control studies designed to test specific hypotheses that are carefully articulated in advance”

Norman Breslow



Overview

1. General introduction to case-control studies
2. Sampling schemes
3. Matching

ESCMID Online Lecture Library
© by author

Aim

- Improved understanding of pros and cons of different types of controls
- Understand different sampling schemes of controls and their implication
- Stimulate to further reading about case-control studies

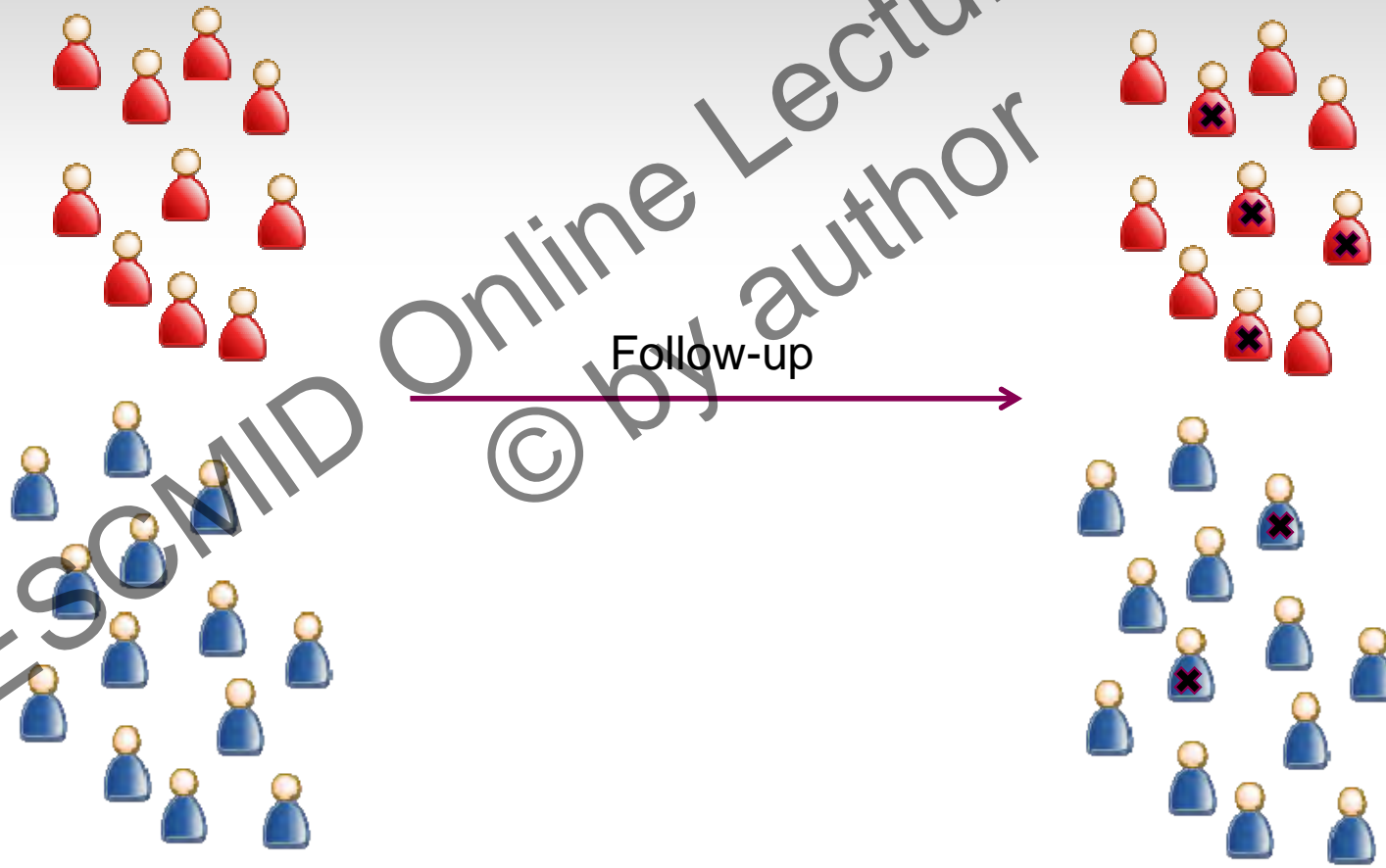


To investigate the attribution of extended spectrum beta-lactamases to mortality in patients with bloodstream infections with *Enterobacteriaceae*?

ESCMID Online Lecture Library
© by author

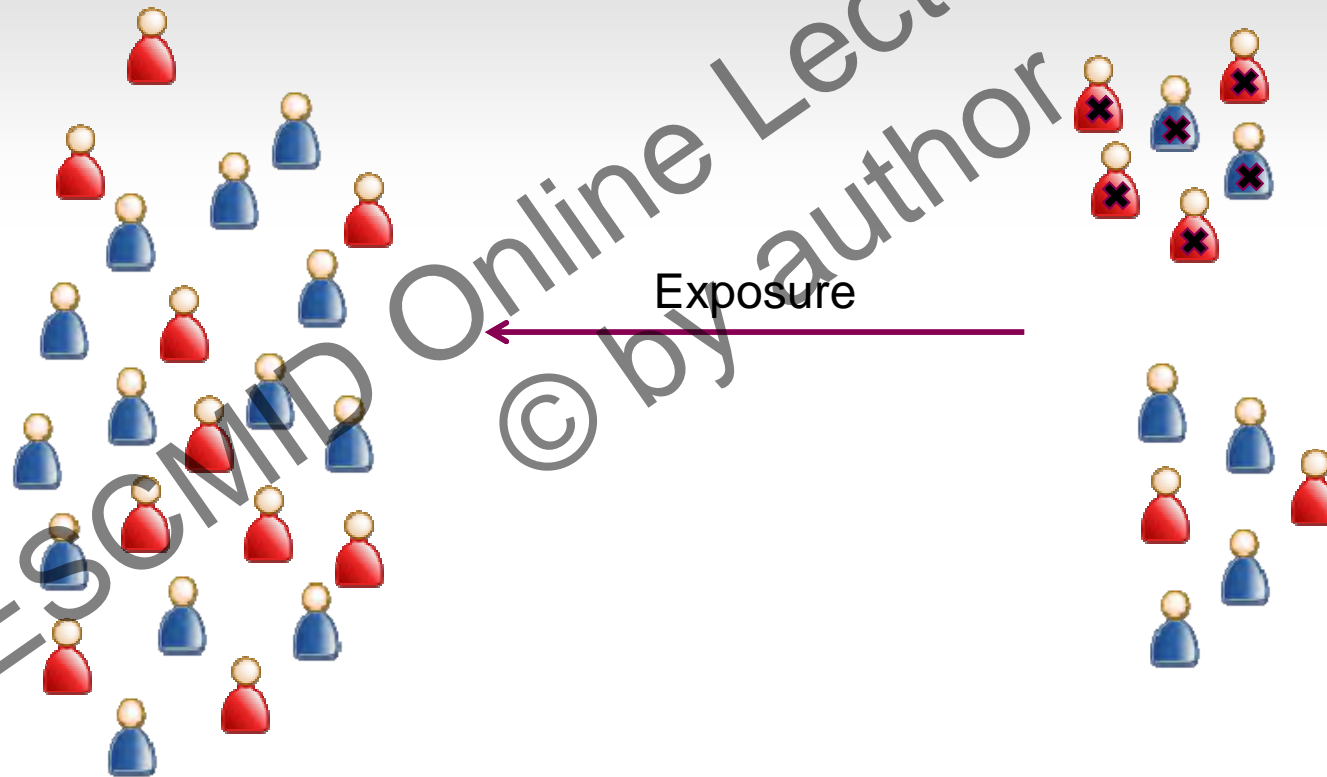


Cohort study



ESCMID Online Lecture Library
© by author

Case-control study



Cohort -> Case-control study

- The goal is to reach the same conclusion from a case-control study as from a cohort study if it had been done
- A case-control study can be conceptualized as a more efficient version of a corresponding cohort study
- It is helpful to first plan the ideal cohort study that would be conducted to investigate the same hypothesis if unlimited resources were available

Advantages of case-control studies

- Efficient
 - Cost
 - Rare diseases
 - Time (long latency/incubation periods)
- Assess a range of potential etiological exposures

Disadvantages of case-control studies

- Susceptibility to bias
 - Selection bias: Selection of controls (and cases)
 - Information bias: non-differential misclassification of exposure status in cases and controls (recall bias)

ESCMID Online Lecture Library
© by author

Fundamentals in control selection

1. Controls should be selected from the same population (or be representable of) - **the source population** - that gives rise to the study cases
2. Because the control group is used to estimate the distribution of exposure in the source population, the cardinal requirement of control selection is that the controls must be **sampled independently of their exposure status**.

Primary vs secondary base

- **Primary base** If the cases are a representative sample of all cases in a precisely defined and identified population and the controls are sampled directly from this source population, the study is said to be *population-based* or a *primary base* study
- **Secondary base** When it is not possible to identify the source population explicitly, simple random sampling is not feasible and other methods of control selection must be used. Such studies are sometimes called studies of *secondary bases*, because the source population is identified secondarily to the definition of a case-finding mechanism. A secondary source population or “secondary base” is therefore a source population that is defined from (secondary to) a given case series. It is often difficult to assess if secondary base is representative of the source population that gives rise to the cases.

1. Are controls selected from a primary or secondary base?
2. Pros and cons of the control selection?

- Research question: What HPV types are associated with cervical cancer in Mozambique?
- HPV types were measured with PCR on cervical samples
- *Cases*: Consecutively enrolled cases at gynaecological department Central hospital, Maputo, Mozambique (referral hospital)
- *Controls*: Women at outpatient clinic of Otolaryngology at the Central Hospital in Maputo

Selection from the primary base

- Advantages:
 - Same study base – i.e. Controls are drawn from the same source population that gave rise to the cases.
 - The distribution of exposures in the controls can be readily extrapolated to the base for purposes such as calculations of attributable risk.
- Disadvantages:
 - Inconvenience – difficult to perform
 - Recall bias. Non-differential misclassification of exposure (e.g. coffee, smoking, solvent)
 - Less motivated than e.g. hospital-based controls to participate.
 - Sometimes difficult to identify cases, e.g. registers

Selection from the secondary base

- Advantages:
 - Convenient
 - Sometimes less risk of information bias
 - Necessary – impossible to determine or sample from primary base

- Disadvantages:
 - It is difficult (impossible?) to assess if controls are representative of the source population that gave rise to the cases.
 - The major problem with any nonrandom sampling of controls is the possibility that they are not selected independently of exposure in the source population. E.g. Smoking

Selection from the secondary base

- Hospital-based controls
 - Referral hospital/clinic, e.g. Mayo Clinic
 - Exclude conditions related to exposure. Usually it is better to select controls from several conditions since it minimized bias if one of the conditions are associated with exposure.
 - Advantages:
 - Likely to be members of the same base as the cases
 - Reduced risk for non-differential exposure misclassification
 - Convenience
 - Disadvantages
 - Distribution of study exposure might be associated with admission to hospital. Hence exposure distribution is not the same as a random sample of the base that gave rise to the cases.
 - Dependence between exposure and condition from where the controls are selected

Selection from the secondary base

- Neighbourhood controls
 - Might not represent source population of cases, e.g. US veterans hospital
 - If neighbourhood is related to exposure, matching need to be taken into account in the analysis
 - Advantages:
 - Does not require a roster
 - Confounding factors associated with neighbourhood may be balanced between cases and controls
 - Disadvantage:
 - Might not satisfy the study base principle
 - Overmatching on the study exposure

Selection from the secondary base

- Random-Digit Dialing
 - Matched on area code, but increasingly difficult with the use of mobile phones
- Relative controls
 - When genetic factors confound the effect of exposure. Risk is overmatching on factors associated with exposures but are not risk factors.

1. Are controls selected from a primary or secondary base?

2. Pros and cons of the control selection?

- Research question: To investigate the cervical cancer risk associated with past HPV exposures (measured with serology).
- A cohort of 13595 women had a health examination at baseline including serum sample
- The cohort was followed for development of cervical cancer using a cancer registry
- Cases: Women who developed cervical cancer during follow-up
- Controls: Selected at random from women without cervical neoplasia at the time when the case was diagnosed. Matching criteria: age, area of residence, type of sample (serum or plasma) and date of enrolment (± 2 months).

Sampling scheme

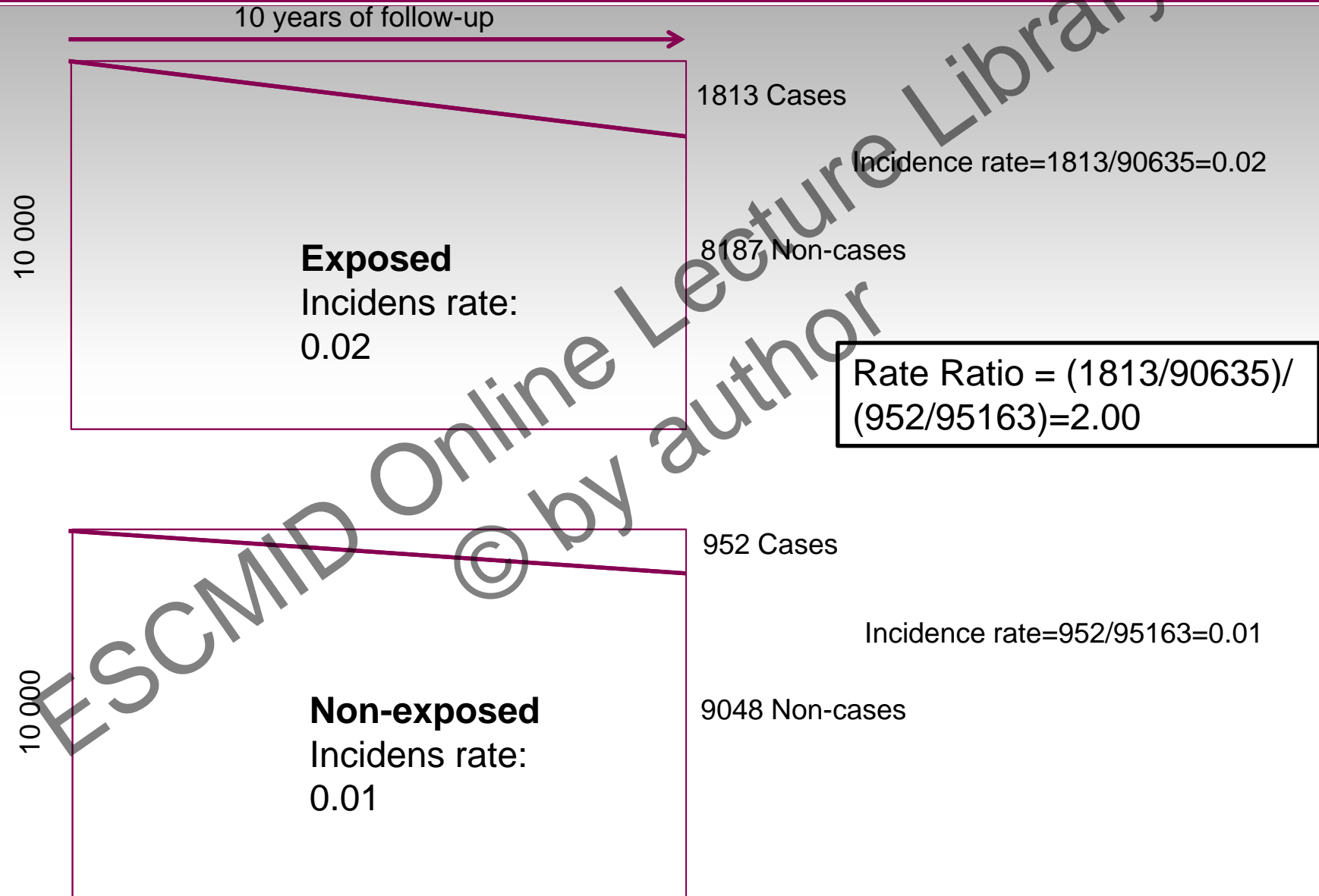
- When controls are selected from members of the population who were at **risk for disease at the beginning of the study's follow-up period**, the case-control **odds ratio estimates the risk ratio** that would be obtained from a cohort design.
- When controls are selected from members of the population who were **non-cases at the times that each case occurs, or otherwise in proportion to the person-time accumulated by the cohort**, the case-control **odds ratio estimates the rate ratio** that would be obtained from a cohort design.
- When controls are selected from members of the population who were **non-cases at the end of the study's follow-up period**, the case-control **odds ratio estimates the incidence odds ratio** that would be obtained from a cohort design.

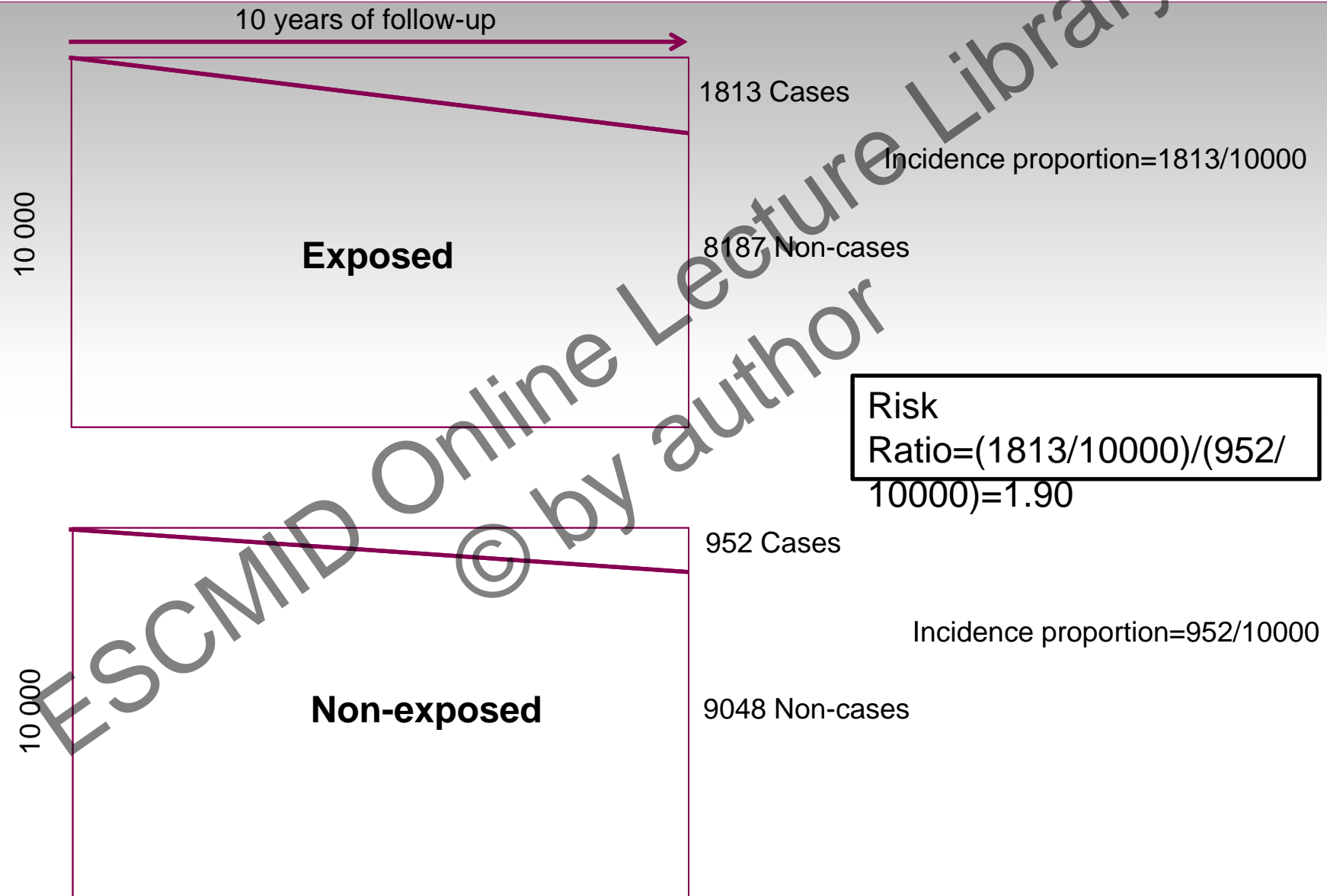
Example

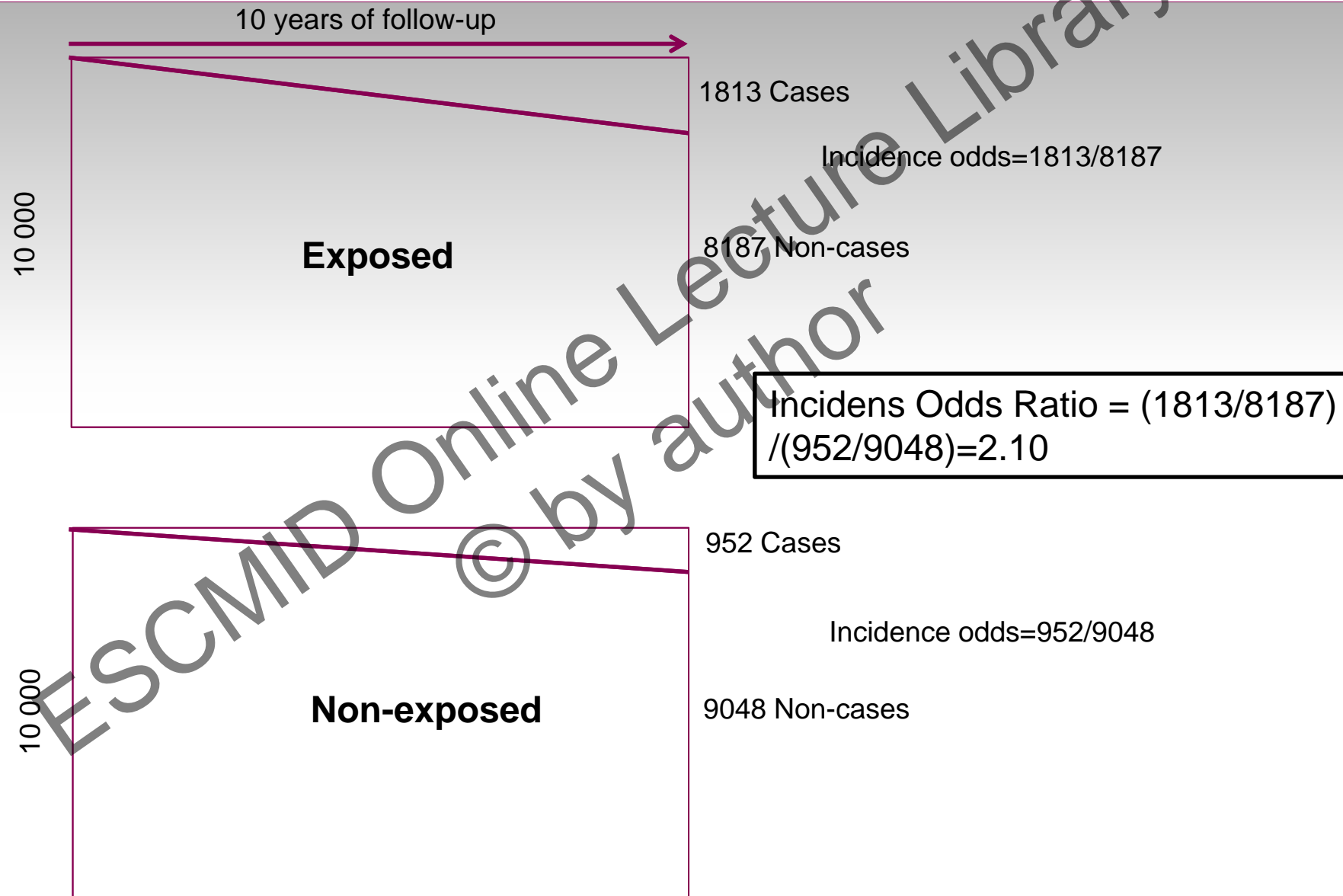
- Two cohorts:

exposed (n=10000), un-exposed (n=10000)

- Incidence rate in exposed: 0.02
- Incidence rate in un-exposed: 0.01
- Follow-up: 10 years





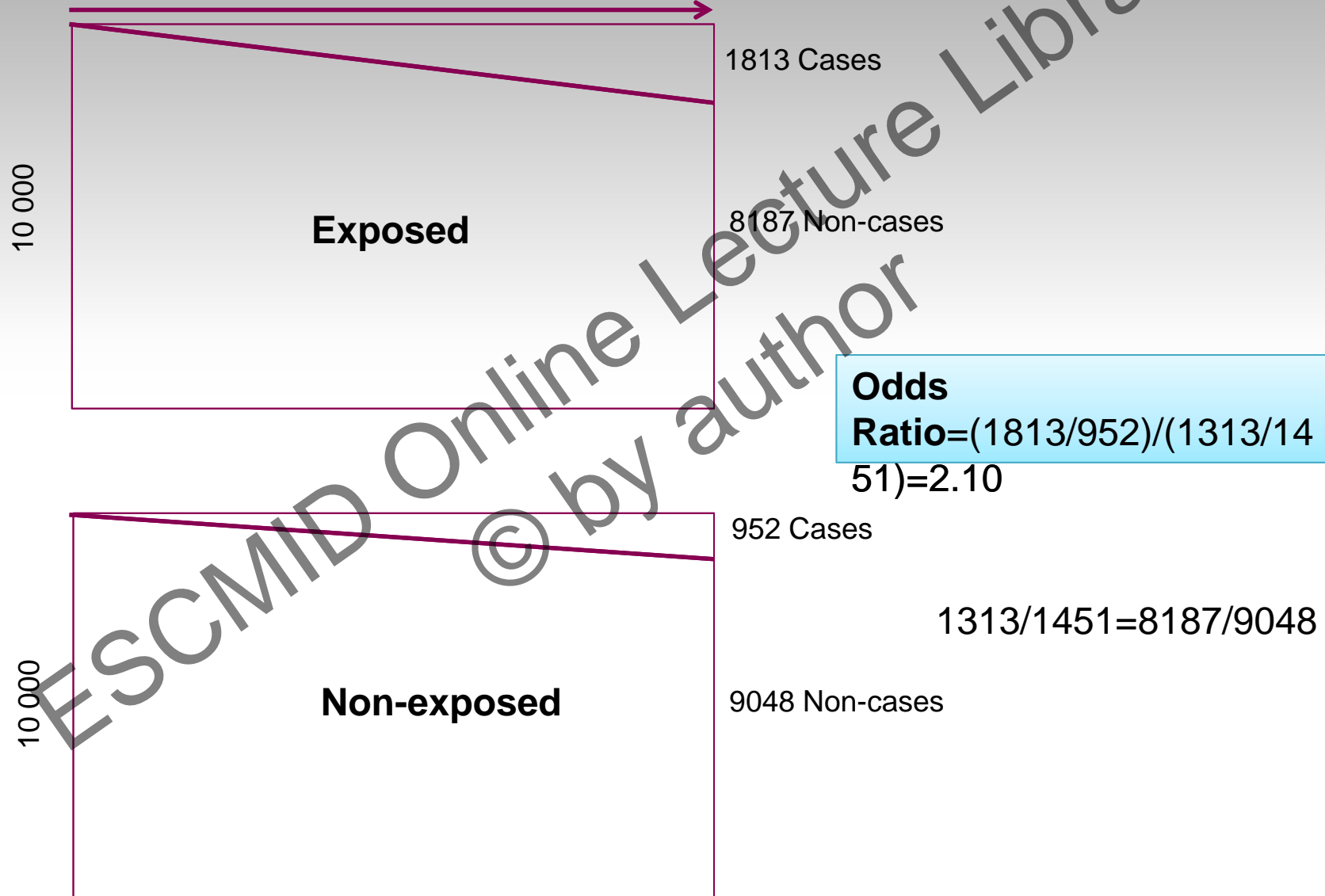




Cumulative case-control

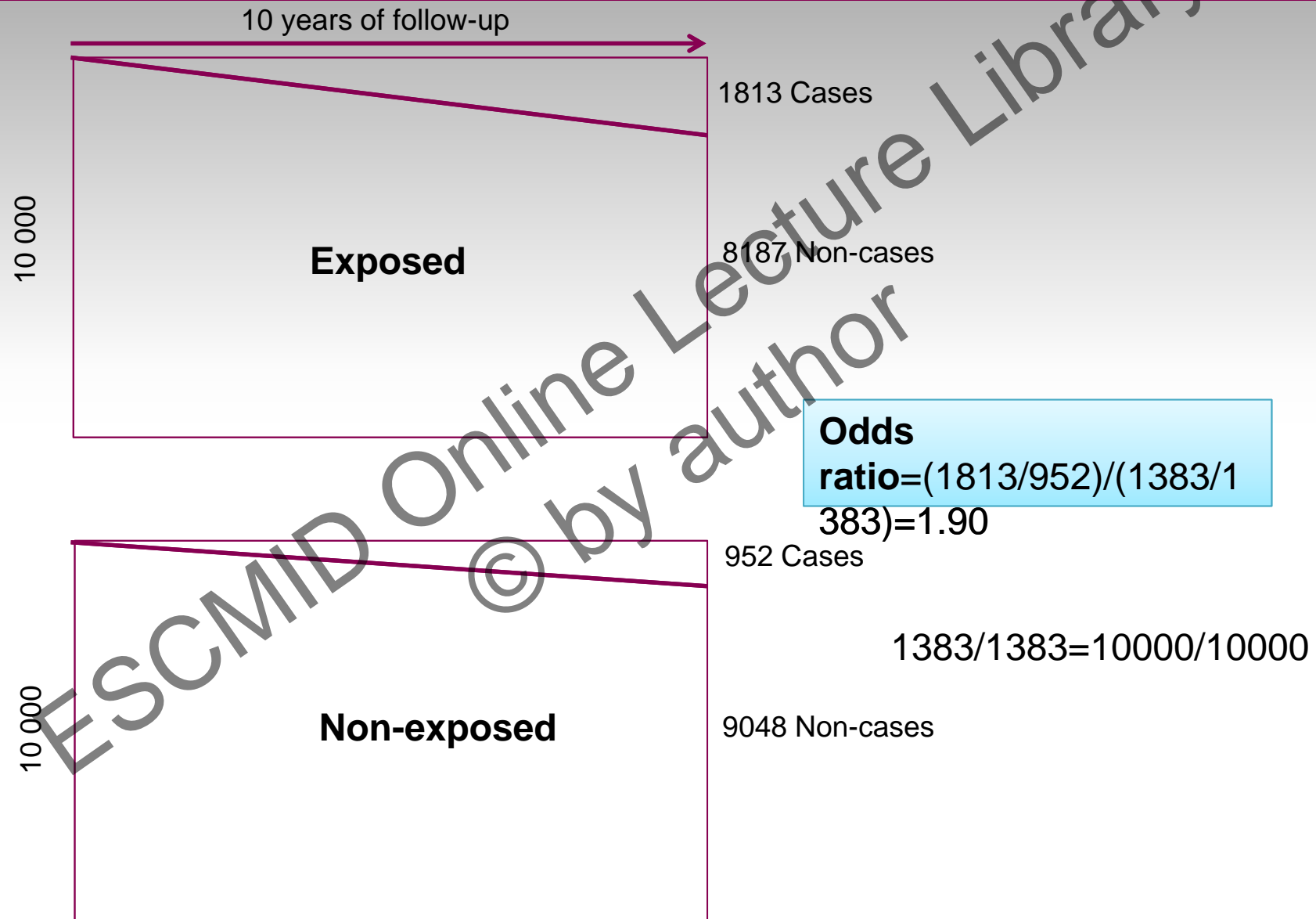
study

10 years of follow-up



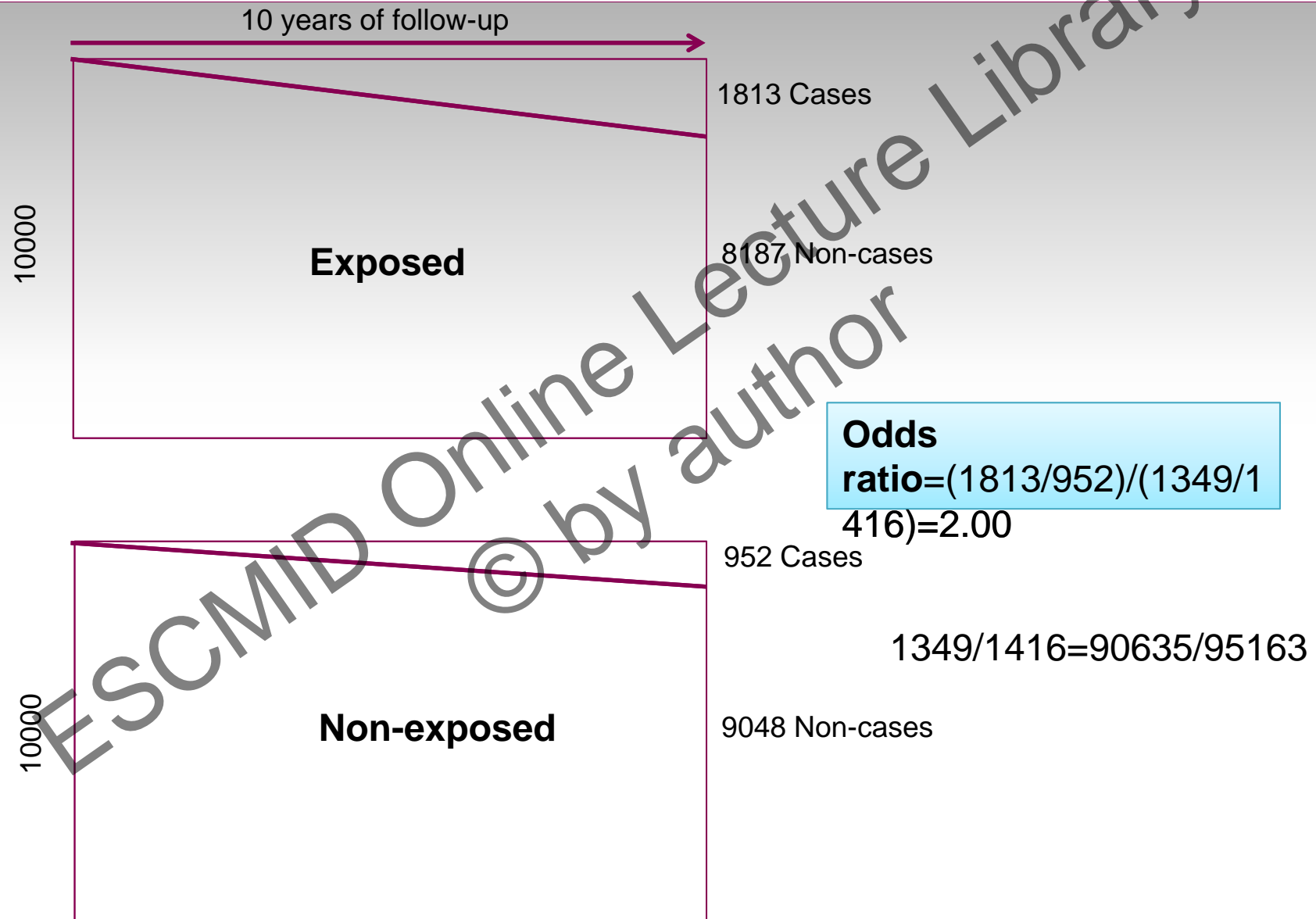


Case-cohort study





Density case-control study



What does this mean in practice?

1. The meaning of the odds ratio is dependent on how controls are being sampled
2. The rarity assumption is dependent on how controls are being sampled
3. Advantages of different sampling schemes

Case-cohort study

Perform several case-control studies using the same control group (if not matched to cases)

Density case-control study

Allows for tight control of the confounding effects of time in the analysis

Cumulative case-control study

Improved statistical efficacy compared to e.g. case-cohort study

1. What sampling scheme was used in this case-control study?

- Research question: To investigate the cervical cancer risk associated with past HPV exposures (measured with serology).
- A cohort of 13595 women had a health examination at baseline including serum sample
- The cohort was followed for development of cervical cancer using a cancer registry
- Cases: Women who developed cervical cancer during follow-up
- Controls: Selected at random from women without cervical neoplasia at the time when the case was diagnosed. Matching criteria: age, area of residence, type of sample (serum or plasma) and date of enrolment (± 2 months).

- 1. Are controls selected from a primary or secondary base?**
- 2. Pros and cons of the control selection?**

- Aim: To investigate the association of viral respiratory infections with pneumonia in children.
- Cases: X-ray verified pneumonia in children <5 years of age seeking healthcare at two paediatric hospitals in Stockholm, Sweden, during 3 years.
- Controls: Consecutively enrolled children at child welfare centers during routine visits (vaccinations) and matched to cases on age and calendar time.
- Nasopharyngeal aspirates were obtained and analyzed by real-time polymerase chain reaction for 15 viruses.

Matching

▪ Advantages

▪ Power

- Can improve efficacy in estimating the effect if the distribution of confounders are substantially different in cases and controls. The gain is usually small!
- Matching can ensure that there are sufficient controls to estimate an effect in a particular subgroup or to identify an interaction.

▪ *Control of unmeasured confounders.* Proxies of unmeasured confounders e.g. Neighbourhood for environmental and socio-economic factors.

▪ *Time comparability.* To achieve time comparability between cases and controls for exposures that vary over time. E.g. epidemic infectious diseases

▪ Disadvantages

▪ *Cost.* Difficulties in obtaining controls

▪ *Exclusion of cases.* When controls are not available.

▪ *Reduced flexibility in analyses.* Can not assess effect of matching variable on outcome.

▪ *Overmatching*

Overmatching

- Matching on a *nonconfounder*. Factor associated with exposure but not disease and vice versa => reduced statistical efficacy
- Matching on an *intermediate variable*. Will bias both unadjusted and adjusted effect estimates (selection bias)

ESCMID Online Lecture Library
© by author

To investigate the attribution of extended spectrum beta-lactamases to mortality in patients with bloodstream infections with *Enterobacteriaceae*?

- Define our cohort: All patients with positive blood culture for *Enterobacteriaceae*
- Select our controls at random from the primary base
- What is T-zero?
- Define mortality (in-hospital, 30-day, 7-day?)
- Sampling scheme? Do we want to estimate attributable proportion?
- Match on age and calendar time (stratified by year?)
- How do we address community-acquired vs. nosocomial infections?
- Should we match on severity of disease? Administered antibiotics? ICU treatment?

Table 3. Effects of method of adjustment on ORs

	No. of analyses	No. of patients	aOR (95% CI)	I ² (%)	P value ^a
All studies ^b	15	7682	1.52 (1.15–2.01)	32	
Adjustments for intermediates					
none	2	398	2.87 (1.57–5.26)	0	0.07
inadequate empirical therapy or sepsis severity	6	1340	2.11 (1.41–3.16)	37	
inadequate empirical therapy and sepsis severity	10	6981	1.39 (1.01–1.92)	30	
Adjustment for inadequate empirical therapy ^c					
no	5	1435	2.77 (2.13–3.60)	0	<0.001
yes	12	7229	1.37 (1.04–1.82)	22	

References

- Modern Epidemiology, K. Rothman
- Selection of controls in case-control studies I-III, S. Wacholder et al, *American Journal of Epidemiology* 1992
- Classification of epidemiological study designs, N. Pierce, *International Journal of Epidemiology* 2012
- Case-control studies N. Breslow in Handbook of Epidemiology by Ahrens and Pigeot
- What Do Case-Control Studies Estimate? Survey of Methods and Assumptions in Published Case-Control Research, Knol et al, *American Journal of Epidemiology* 2008
- Literature about case-control studies of antibiotic resistance
 - Methodological Principles of Case-Control Studies That Analyzed Risk Factors for Antibiotic Resistance: A Systematic Review, Harris et al *Clinical Infectious Diseases* 2001
 - Impact of severity of illness bias and control group misclassification bias in case-control studies of antimicrobial-resistant organisms, Harris et al, *Infection Control and Hospital Epidemiology* 2005
 - The case-case-control study design: addressing the limitations of risk factor studies for antimicrobial resistance, Kaye et al, *Infection Control and Hospital Epidemiology* 2005
 - Epidemiological interpretation of studies examining the effect of antibiotic usage on resistance, Schwaber et al, *Clinical Microbiology Review* 2013



Karolinska
Institutet

ESCMI Online Lecture Library
© by author

Thank you!